 Human Molecular Genetics, 2002, Vol. 11, No. 20 2417–2423

The allelic architecture of human disease genes: common disease–common variant . . . or not?

Jonathan K. Pritchard* and Nancy J. Cox

Department of Human Genetics, University of Chicago, 920 E 58th St–CLSC 507, Chicago IL 60637, USA

Received July 16, 2002; Accepted July 24, 2002

Linkage disequilibrium (LD) plays a central role in current and proposed methods for mapping complex disease genes. LD-based methods work best when there is a single susceptibility allele at any given disease locus, and generally perform very poorly if there is substantial allelic heterogeneity. The extent of allelic heterogeneity at typical complex disease loci is not yet known, but predictions about allelic heterogeneity have important implications for the design of future mapping studies, including the proposed genome-wide association studies. In this article, we review the available data and models relating to the number and frequencies of susceptibility alleles at complex disease loci—the 'allelic architecture' of human disease genes. We also show that the predicted frequency spectrum of disease variants at a gene depends crucially on the method of ascertainment, for example from prior linkage scans or from surveys of functional candidate loci.

One of the central challenges in modern human genetics is to unravel the genetic basis of human diseases and other phenotypes of interest. Apart from the inherent interest in understanding the biological determinants of phenotypic variation, it is hoped that this work will lead to important medical advances (1). Most notably, determination of the genetic variants involved in a particular disease should provide new insight into the disease etiology, may suggest novel pharmaceutical targets, and could lead to genetic screening to identify individuals at increased risk.

The genetics community has achieved great success in finding the genes that are responsible for a wide range of Mendelian diseases (2). In contrast, the search for complex disease genes has been relatively frustrating (3), despite intense research effort in both the academic and commercial sectors. Linkage mapping, which is a powerful tool for finding Mendelian disease genes, often produces weak, and sometimes inconsistent, signals in complex disease studies (4). To date, only a few variants that contribute to complex diseases have been conclusively identified.

At the present time, little is known about the determinants of complex diseases. For a given disease, an individual's risk probably depends on some unknown function of genetic, environmental/lifestyle and stochastic factors. For some complex diseases, there are closely related monogenic forms [e.g. *MODY* genes for diabetes (5)], but typically it seems that no single factor is either necessary or sufficient for disease. This, in essence, is the challenge of complex disease mapping: the marginal increase in risk due to the at-risk genotype at a disease gene is quite small. To succeed in finding complex disease genes, a study must detect a relatively weak statistical signal, and choices in study design can potentially have a dramatic impact on the probability of success (6).

The focus of this article is on the allelic architecture of complex disease genes, because this is a critical factor in determining which study designs will be most successful. By 'allelic architecture', we refer to the number of distinct alleles that impact disease susceptibility at a given disease locus, and their frequencies. A full description of allelic architecture would also include the penetrances of all the genotype combinations, but little is known about this. Instead, we describe predictions for a rather simple model with two equivalence classes of alleles: 'normal' (N) and 'susceptibility' (S) alleles, where S alleles lead to increased disease risk (7). We focus mainly on two quantities that are especially relevant for study design: (i) the total frequency of S alleles and (ii) the degree of allelic identity (8) (inversely related to the number of alleles) within the S class. Both of these quantities are actually random values that depend on the outcome of the evolutionary processes of mutation, selection and random drift, so it is most accurate to talk about the *distribution* of outcomes (7,9).

A popular hypothesis about allelic architecture proposes that most of the genetic risk for common, complex diseases is due to disease loci where there is one common variant (or a small number of them) (8,10,11). If true, the 'common disease–common variant' (CDCV) hypothesis implies that association mapping should be a powerful tool for detecting complex disease loci of small effect, and provides an important impetus for the Haplotype Map Project, and the proposed genome-wide association scans of the future (6,12). Alternatively, if there is

*To whom correspondence should be addressed. Tel: +1 7738345248; Fax: +1 7738340505; Email: pritch@uchicago.edu

substantial allelic heterogeneity (low allelic identity) at each locus, association mapping will be much more difficult, because each new mutation arises on an independent haplotype background, and the different mutations tend to cancel out each other's signals (13).

This article provides an overview of current information about the allelic architecture of human disease genes, based on both empirical data and theoretical models. We also discuss some of the implications for disease gene mapping.

EXAMPLES FROM KNOWN DISEASE GENES

Mendelian disorders

There are now extensive data on the allelic architecture of Mendelian disease genes (2,8,14). For Mendelian diseases, the total frequency of susceptibility mutations is usually very low (usually $\ll 1\%$). For the vast majority of Mendelian diseases, the overall mutation frequency is probably determined by mutation–selection balance, i.e. by a balance between the input of new mutations and purifying selection that removes them (8,15).

Mendelian disorders often feature extremely high levels of allelic heterogeneity. For example, a study of 424 UK families with Haemophilia B found 167 distinct mutations, and concluded that there had been at least 302 independent mutation events (16). The most frequent mutation thought to derive from a single common ancestor was present in just 5% of the families. In contrast, cystic fibrosis (CF) has one major allele, ΔF508, that is relatively common, accounting for 67% of European CF chromosomes. There are also many rare CF alleles; one large study identified 272 different mutations among 27 000 CF chromosomes (17). Overall, there is a strong trend towards decreased allelic heterogeneity as the overall frequency of disease mutations increases (8); we return to this observation later in the article.

A small number of Mendelian mutations are at frequencies $>1\%$ in the general population. For several of these, it has been proposed that the higher frequencies may be the result of past selection in favour of heterozygotes. For example, β-globin defects that lead to sickle cell anaemia protect heterozygotes against malaria (15). More controversially, it has been suggested that carriers of the ΔF508 mutation enjoyed a selective advantage in past epidemics of cholera (18,19).

Complex traits

In contrast to Mendelian diseases, much less is known about the properties of complex disease genes. Thus far, relatively few susceptibility variants for complex diseases have been unambiguously identified. Moreover, the few variants that have already been found may not be representative of complex disease variants in general. It is quite likely that these early successes are 'low-hanging fruit'—that is, the susceptibility variants are probably more penetrant, and may have simpler allelic architecture than most complex disease genes. Nonetheless, it is of interest to review the properties of some of the loci found so far; the following list is not exhaustive, but provides some indication of the diversity of allelic architectures.

A favourite example used by the CDCV proponents is the *APOE* locus, where a single common allele (known as ε4) increases risk to Alzheimer disease and heart disease (20). The ε4 allele is at frequencies in the range 0.05–0.41 in various populations (21). The *PPARγ* locus, implicated in type 2 diabetes, also has a single major susceptibility variant (the Pro12Ala polymorphism), but this is at very high frequency (0.85) (22), as is the tandem repeat variation (VNTR) at the INS locus that is associated with type 1 diabetes (frequency 0.75) (23). The allelic architecture at the *NOD2* locus, involved in Crohn's disease, is rather more complicated, including several nonsynonymous mutations and a frameshift mutation, each at frequencies $\geq 1\%$ in controls, and with a total frequency of susceptibility mutations of $\sim 10\%$ in the general population (24,25). In contrast to *APOE*, *NOD2* and *PPARγ*, the apparent role of the *CAPN10* locus in type 2 diabetes (26) cannot be explained by non-synonymous mutations. Instead, the mechanism seems to be strikingly complex, with individuals who carry a particular *pair* of discordant haplotypes showing significantly increased risk. The two haplotypes are defined by combinations of three non-coding single-nucleotide polymorphisms (SNPs), and are at frequencies of 0.32 and 0.23, respectively, in the study population of Mexican Americans.

In summary, the available sample of known complex disease genes is too small, and possibly too biased, to draw general conclusions about the allelic architecture of complex trait genes, but the available examples are interesting for the diversity of architectures that they represent.

One additional observation that may be relevant is the fact that positional cloning of the genes responsible for linkage signals has proved so challenging up to now. This is despite the efforts of many groups, working on many different diseases. A possible explanation is that most genes are hard to find because they have complex allelic architecture, but we cannot exclude the possibility that these studies have low power for other reasons.

Impact of selection

A curious feature of the *APOE* locus is that the ε4 allele that is associated with increased risk for various diseases is actually the ancestral allele (21). Similarly, we noted above that the *PPARγ* and *INS* variants that increase susceptibility to disease are at very high frequency. Again, we should be cautious about the impact of the ascertainment bias described above, but these observations seem surprising under a model where the susceptibility alleles are purely deleterious, as one might imagine for disease-susceptibility alleles. [For a discussion of selection in late-onset diseases, see (7).] An alternative hypothesis is that variants involved in certain modern diseases may have been subject to frequency-dependent selection, or fluctuating selection in the past [as with the 'thrifty genotype' hypothesis for diabetes (27), and the β-globin and CF examples described above]. In the next section, we focus on population genetic models of allelic architecture. The available work focuses on models of strictly deleterious variation; however, the

latter observations suggest that other types of selection may also be important.

THEORETICAL MODELS

It is possible to gain further insight into the likely allelic architecture of complex diseases using theoretical models (7–9,28). The basic idea is to construct simple models of the genetics of a disease, as well as the evolutionary processes of selection, mutation and genetic drift. Analysis of these models should provide insight into the likely properties of the allelic architecture of complex disease loci. Clearly, such models are a simplification of the complexities of real life, but the hope is that the models capture the most salient features for most loci.

We start by assuming that in the genome there are L genes that, if mutated, could contribute to susceptibility for the disease of interest. At each locus, there are two possible classes of alleles: 'normal' (N) and 'susceptibility' (S) alleles (7,9). There is a mutation rate from N to S alleles, and a (much lower) reverse mutation rate. There is also the possibility of weak purifying selection against S alleles. Pritchard (7) provides further discussion of these modelling choices and their implications.

For population genetic modelling, it is also necessary to make some assumptions about population history. The simplest model assumes a random-mating population with a constant effective size N_e (usually estimated to be $\sim$10 000 for humans). This model ignores various complexities, including, notably, the dramatic increase in human population size to present numbers, which is predicted to impact the allele frequency distribution (8). The timing and magnitude of the population growth is somewhat controversial (8,29–35); however, it seems that the impact of growth may be modest enough that, apart from low frequency variants (discussed below), the constant population size model may provide an adequate approximation to the overall frequency spectrum (33).

There are three main parameters in this model: the mutation rate to susceptibility alleles, per gene, per generation (μ_S), the reverse mutation rate (μ_N), and the strength of purifying selection against S alleles (s) (7,9). It turns out that the mutation rate to disease alleles (μ_S) is a critical parameter in determining the allelic architecture. In particular, higher mutation rates lead to greater allelic heterogeneity. Two approaches have been used to predict μ_S for complex disease loci. The first is to assume that mutation rates at Mendelian loci are typical of those at complex disease loci (7,8). This assumption could lead to underestimates, since there may be more sites in a complex disease gene that can produce low-penetrance mutations than sites in a Mendelian gene that can produce highly penetrant Mendelian mutations. The second approach is to estimate the number of sites in a typical gene that could mutate to produce susceptibility mutations, and then multiply by a per-site mutation rate (7); this obviously involves considerable uncertainty as well. Reich and Lander (8) estimated a range of mutation rates of 10^{-7}–10^{-4}, and used the geometric mean of 3.2×10^{-6} in their analysis. Pritchard (7) estimated a range of 2.5×10^{-6}–1.3×10^{-4}, but argued that the upper part of this range was most relevant.

In population genetics, it is conventional to rescale the mutation and selection rates by $4N_e$ because this factor appears in all the mathematics (15,36); hence we define the scaled forward and reverse mutation rates $\beta_S = 4N_e\mu_S$ and $\beta_N = 4N_e\mu_N$, and selection rate $\sigma = 4N_e s$. The Reich and Lander (8) estimate of β_S is 0.13 (prior to population expansion), and the Pritchard (7) range is 0.1–5.0.

Total frequency of susceptibility mutations

When there is strong selection, as for Mendelian diseases, the overall frequency of susceptibility mutations, p, is determined by a balance between mutation and selection. Mutation creates new alleles, while selection acts to remove them. The equilibrium frequency is $p = \sqrt{\beta_S/\sigma}$ for recessive mutations and $p = \beta_S/\sigma$ for dominant mutations (15). Notice that this equilibrium for p does not depend on population size; it assumes that selection is strong, and stochastic effects including drift are ignored.

Mutation–selection balance is only an appropriate model for genes where the equilibrium value of p is quite low [e.g., $<1\%$; but see ref. (8)]. If the equilibrium for p is assumed to be large then, for plausible mutation rates, this implies that the strength of selection is very weak. With weak selection, the stochastic effects of mutation and genetic drift start to dominate and as described below, most loci will not actually be close to the deterministic equilibrium.

When selection is weak, as seems likely for many complex disease mutations, genetic drift becomes important, and the total frequency of susceptibility mutations, p, is expected to vary widely among loci. Under this model, the probability distribution for p can be written down in a simple form, first obtained by Sewall Wright (7,36,37). The red lines in Figure 1 show examples of this frequency distribution for three sets of parameter values. In the examples, the forward mutation rate β_S takes the suggested values of 0.1 (8), and 1.0 (7). The reverse mutation rate, β_N, is set substantially lower (0.01), as expected on biological grounds (7).

These lines illustrate several important points. First, for most parameter values, the vast majority of the probability mass is on p near 0 (or sometimes near 1, depending on σ) (7). To put this another way: there may be many genes in the genome that *could* impact disease risk, if mutated. However, the vast majority of these are likely to have very low variability, and hence they contribute little to the inherited variation in risk. Using parameter values that were loosely based on autism, Pritchard (7) found that for a particular model with 100 loci, the vast majority of the variation in genetic risk was due to just 5 loci; these are the loci that happen, by chance, to be more variable than average. The model predicts that most loci that are good biological candidates will not have S alleles at intermediate frequencies. This prediction seems to be reasonably robust unless (i) the mutation rate β_S is higher than expected or (ii) selection favouring heterozygotes is surprisingly common (7).

A second point is that mutation rates and the strength of selection are likely to vary substantially among loci. Recall that the mutation rate that matters here, β_S, is the total rate of mutation to S alleles for the entire gene. β_S is sure to vary widely, since different genes will have different numbers of sites that can mutate to produce susceptibility alleles. Genes with high mutation rates and weak purifying selection are

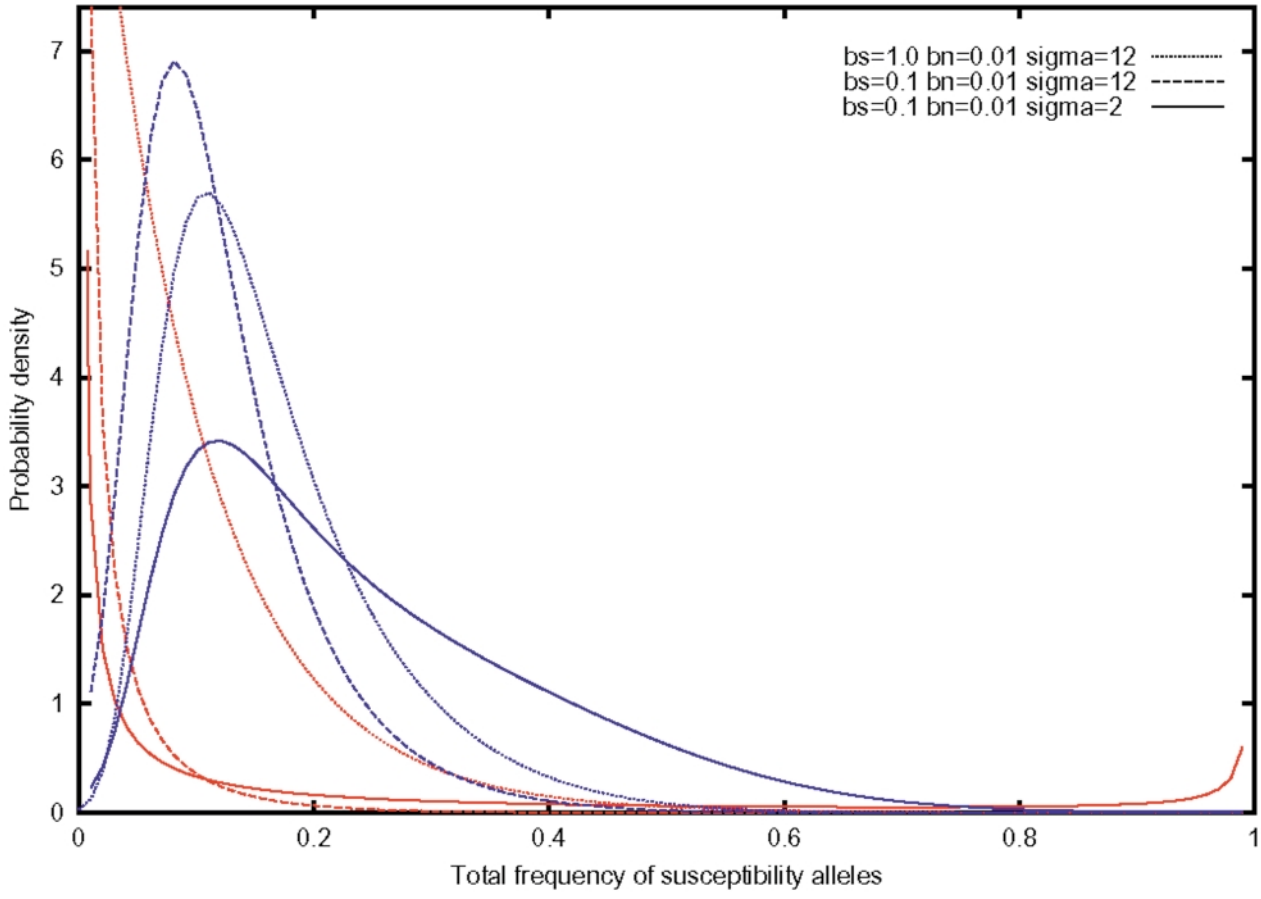

Figure 1. The distribution of the total frequency of susceptibility mutations: (red) at a random disease locus and (blue) at a disease locus that has been detected by a linkage scan. Each line shows the probability density of the total frequency of S alleles (i.e. the probability that the allele frequency lies between frequencies p_1 and p_2 is the area under the curve between p_1 and p_2). The three pairs of lines correspond to different mutation rates and levels of purifying selection. The red lines show the background distribution of allele frequencies at random disease loci. Unless the mutation rate β_S is very high, allele frequencies at most random disease loci are near 0 (or sometimes near 1) under this type of model. But when using linkage mapping to find low-penetrance alleles, we only have power to detect loci with intermediate frequencies. The blue lines show the posterior distributions of allele frequencies given that the locus was detected using affected sib pairs. Notice that these distributions are substantially shifted towards higher frequencies. The plots assume the model of Pritchard (7), with forward and reverse mutation rates β_S and β_N and strength of selection σ shown in the key (*bs*, *bn* and *sigma*, respectively). The blue lines were computed using that model as a prior in combination with the model and methods of Risch and Merikangas (6), with a multiplicative risk factor $\gamma = 3.0$ per allele, and assuming that a P-value of 10^{-4} was obtained using 300 affected sib pairs with perfect inheritance information.

considerably more likely to have S alleles at intermediate frequency, and hence the loci that contribute most to population risk will tend to have higher mutation rates than average (7).

Lastly, we point out that standard mapping techniques, such as linkage mapping and association mapping, only have power to detect low-penetrance alleles at loci where the class of S alleles is at intermediate frequency (6). Thus, even though most loci may have very low frequencies of S alleles, the frequency distribution at loci that we actually find by linkage mapping is strongly skewed towards intermediate frequencies (blue lines in Fig. 1), compared with the background distribution (red lines).

Allelic heterogeneity of susceptibility mutations

We now turn our attention to the distribution of S alleles within the susceptible class. This does not affect linkage mapping, but is important for association mapping, where allelic heterogeneity can be a severe problem (13).

Classical results from population genetics provide the distribution of frequencies of neutral alleles (15,36). The probability that two alleles, chosen at random from a population of constant size, are identical (allelic identity) is

$$\frac{1}{1+4N_e\mu},$$

where μ is the total mutation rate to new alleles. The expected number of distinct alleles in a random sample of n chromosomes is

$$\sum_{i=0}^{n-1}\frac{4N_e\mu}{4N_e\mu+i}$$

The complete distribution of the allele counts is given by the so-called Ewens sampling formula, and is also a function of $4N_e\mu$ (36).

A remarkable result is that these same expressions, originally derived for the entire population, hold for the allelic distribution *within* the susceptible class, after substituting $\beta_S(1\text{-}p)$ in place of $4N_e\mu$ (8,9,28). The result holds provided that all *S* alleles are selectively equivalent.

It is clear from these expressions that β_S plays a central role in determining the degree of allelic heterogeneity. If β_S is small (0.1, say), then the predicted allelic identity is very high. For β_S in the upper part of the predicted range (from 1 to 5), the expected allelic identity is considerably lower. Another way to look at this is to ask what proportion of the total class of *S* alleles is made up by the single most common allele. For $\beta_S = 0.1$, the most common *S* allele represents ~94% of all susceptibility alleles, on average. This drops to 65% and 33% as β_S increases to 1.0 and 5.0, respectively (7). As described above, the upper part of this range may be most relevant, because such loci are much more likely to be polymorphic. In summary, it is predicted that at many loci the levels of allelic heterogeneity may be high enough to be a non-trivial impediment to association mapping. However, under the model of constant population size, *extremely* high levels of heterogeneity are not expected, unless β_S is surprisingly high.

But the theory described above for constant population size does not work well for the allelic distribution at Mendelian loci, where (i) there is often extreme allelic heterogeneity, as described above for haemophilia B (16), (ii) the sampling properties and ages of susceptibility alleles often do not fit neutral expectations (28,38) and (iii) the level of allelic heterogeneity is strongly inversely related to the overall frequency of *S* alleles (8). Reich and Lander (8) argued that these findings can be explained by the impact of population growth. In effect, population growth increases β_S, so that β_S is perhaps 10^5 times higher today than the long-term average value of $\beta_S \approx 1$. The time-scale for changes in the level of allelic heterogeneity depends on the total *S*-allele frequency, *p*. Loci where *p* is very small adjust to the increased population size quite quickly, while loci with higher *p* (e.g. >0.01), adjust much more slowly (8). This means that while the recent population growth seems to have had a dramatic impact on the frequency distribution for very rare Mendelian mutations, greatly increasing the extent of allelic heterogeneity, it has probably had little impact for loci with higher frequencies of *S* alleles (8).

LINKAGE MAPPING VERSUS FUNCTIONAL CANDIDATES

The results described in the previous section lead to a crucial point: *a gene association study that begins by looking at* a priori *functional candidates is looking, on average, for a different class of variants than is a positional cloning study that started from a significant linkage signal*.

The theoretical results imply that at most disease loci, we expect the cumulative frequency of all contributing alleles at a particular locus to be in the very low or very high range (near 0 or near 1). Overall, each of these loci will contribute little to the population risk for disease. Only a relatively small subset of disease loci are likely to have susceptibility alleles at intermediate frequencies. However, these loci are likely to contribute much of the total variance in risk, and hence they are the loci that could possibly be detected by linkage studies (Fig. 1).

Conversely, when a study focuses on functional candidates, the probability that a given locus will have variants at intermediate frequencies is low. It is far more likely that the variation will be rare, and, if so, it may be as heterogeneous as many Mendelian diseases, for the reasons described above (8). Studies should be designed accordingly. Genotyping a small number of common SNPs is likely to have low power, unless the candidate happens to be one of the rare genes with common variants. In particular, if the candidate locus lies in a region that has not previously produced suggestive LOD scores, this further argues against the presence of common variants, unless they have quite low penetrance, in which case large sample sizes will be required.

The results of diabetes studies carried out thus far are reasonably consistent with these expectations. While many different polymorphisms across the entire spectrum of allele frequency have at one time or another been reported to show association with diabetes and related phenotypes, only a couple of these are widely accepted as affecting susceptibility to diabetes, both identified through candidate gene studies. As noted above, the Pro12Ala polymorphism at the *PPARγ* locus has a rather high frequency allele increasing susceptibility to type 2 diabetes (22). This variation would not generate evidence for linkage using typical sample sizes, but rather was identified through candidate gene studies. Similarly, the variation at the *INS* VNTR that is associated with susceptibility to type 1 diabetes has relatively high frequency in populations of European descent and was discovered through candidate gene studies (23,39). Evidence for linkage was negligible at *INS* until more recent studies were conducted on very large numbers of families (>700) (40). The only variation for type 2 diabetes thus far reported to be identified through a positional cloning study (based originally on a linkage signal) is at *CAPN10* (26,40). The variation at *CAPN10* hypothesized to increase risk of type 2 diabetes does occur at intermediate frequency in the population in which substantial evidence for linkage in this region was observed.

DISCUSSION

As we noted at the start of this article, the allelic architecture of complex diseases is a critical unknown in the design of gene mapping studies of the future, including, in particular, the proposed genome-wide association studies. Many Mendelian disorders display extreme levels of allelic heterogeneity that would doom LD-based techniques for finding low-penetrance complex disease variants. The CDCV hypothesis represents the best-case scenario for large-scale association mapping, so it is important to ask whether this hypothesis is unreasonably optimistic. In particular, why should the architecture of complex disease loci differ from that of Mendelian loci?

Mendelian disease mutations are highly penetrant, and usually under very strong selection, which keeps them at low frequencies. Susceptibility variants involved in complex diseases seem to have low or medium penetrance, and are

probably not subject to such strong selection. This suggests one crucial difference: it is possible for the class of *S* alleles at a complex disease locus to reach intermediate frequencies (5–10% or more). Such loci are likely to represent only a small fraction of all the loci that are involved in disease susceptibility, but they will contribute disproportionately to the total population risk. They are also the loci that we can hope to find using linkage or association scans. Thus, the class of genes that we are looking for in the next generation of gene mapping studies probably has a much higher total frequency of susceptibility alleles than do most Mendelian loci. So what does this mean for the extent of allelic heterogeneity?

It is interesting to compare the conclusions of the two studies that have looked most directly at this question—those by Pritchard (7) and by Reich and Lander (8). Apart from technical differences (Reich and Lander used a deterministic model for *p*, the total frequency of *S* alleles, while Pritchard used a stochastic model), the major modelling differences were that Reich and Lander used a lower mutation rate and incorporated population growth. As noted above, Pritchard favoured the higher mutation rate for complex disease loci in part because such loci are substantially more likely to be polymorphic. Modelling population growth appears to lead to a considerably better description of very low-frequency alleles, but it makes less difference for intermediate values of *p* (8). At intermediate *p*, the population growth assumption has an effect similar to reducing the difference in mutation rate between the two studies. In the end, the numerical predictions about allelic heterogeneity for intermediate *p* are not very different in the two studies. But the conclusions differ in tone: Reich and Lander described allelic identity above 0.1 as being a 'simple' allelic spectrum (compared with Mendelian diseases, presumably), even though this indicates the presence of many different alleles, none at high frequency. For loci with the highest mutation rates, Pritchard noted that 'it is unlikely that any single mutation will constitute a large fraction of the susceptible class. In this case association mapping is not very powerful'.

In summary, the theoretical predictions are somewhat, but not entirely, encouraging. If current mutation rate predictions for complex disease loci are in the right range (7,8), then it seems that loci with intermediate *p* are not expected to have devastating levels of allelic heterogeneity. But the levels of heterogeneity are still likely to be problematic in many cases. The current conception of association mapping one marker at a time (6) or one 'haplotype block' at a time (12) will perform poorly if the allelic identity is as low as 0.1. But allelic associations with particular susceptibility variants should extend over reasonably large distances (7), and it is plausible that when there are multiple alleles, more sophisticated statistical methods could extract considerably more information from multilocus genotype data. From the viewpoint of experimental design, it seems that positional cloning studies and, to a greater extent, functional candidate studies, need to perform extensive sequencing, in large numbers of individuals, in case the variation that they are searching for consists of low-frequency variants.

The *NOD2* example provides a good cautionary tale for would-be mappers of complex disease genes (24,25). Several features contributed to the success in finding this gene despite the presence of moderate allelic heterogeneity: the mutations had relatively high penetrance (24), the gene made good biological sense, and the susceptibility variants included a frameshift mutation and several non-synonymous mutations. But would our current mapping strategies be able to find a gene with similar allelic architecture under less favourable circumstances—for example, with lower penetrance, in a gene of unknown function, with multiple non-coding variants?

ACKNOWLEDGEMENT

J.K.P. was supported in part by a Hitchings Elion fellowship from the Burroughs Wellcome Fund.

REFERENCES

1. Collins, F.S. and Guttmacher, A.E. (2001) Genetics moves into the medical mainstream. *JAMA*, **286**, 2322–2324.
2. Cooper, D.N., Ball, E.V. and Krawczak, M. (1998) The human gene mutation database. *Nucleic Acids Res.*, **26**, 285–287.
3. Risch, N. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
4. Altmüller, J., Palmer, L.J., Fischer, G., Scherb, H. and Wjst, M. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.*, **69**, 936–950.
5. Fajans, S.S., Bell, G.I. and Polonsky, K.S. (2001) Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N. Engl. J. Med.*, **345**, 971–980.
6. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
7. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to common diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
8. Reich, D.E. and Lander, E.S. (2001) On the allelic spectrum of human disease. *Trends Genet.*, **17**, 502–510.
9. Hartl, D.L. and Campbell, R.B. (1982) Allele multiplicity in simple Mendelian disorders. *Am. J. Hum. Genet.*, **34**, 866–873.
10. Lander, E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.
11. Chakravarti, A. (1999) Population genetics—making sense out of sequence. *Nat. Genet.*, **21**, 56–60.
12. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
13. Slager, S.L., Huang, J. and Vieland, V.J. (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet. Epidemiol.*, **18**, 143–156.
14. Wright, A.F., Carothers, A.D. and Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nat. Genet.*, **23**, 397–404.
15. Hartl, D.L. and Clark, A.G. (2000) *Principles of Population Genetics*. Sinauer, Sunderland, MA.
16. Green, P.M., Saad, S., Lewis, C.M. and Giannelli, F. (1999) Mutation rates in humans. I. Overall and sex-specific rates obtained from a population study of hemophilia B. *Am. J. Hum. Genet.*, **65**, 1572–1579.
17. Estivill, X., Bancells, C. and Ramos, C. (1997) Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. *Hum. Mutat.*, **10**, 135–154.
18. Gabriel, S.E., Brigman, K.N., Koller, B.H., Boucher, R.C. and Stutts, M.J. (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science*, **266**, 107–109.
19. Wiuf, C. (2001) Do ΔF508 heterozygotes have a selective advantage? *Genet. Res.*, **78**, 41–47.
20. Corder, E.H., Saunders, A.M., Strittmatter, W.J., Schmechel, D.E., Gaskell, P.C., Small, G.W., Roses, A.D. Haines, J.L. and Pericak-Vance, M.A. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, **261**, 921–923.

21. Fullerton, S.M., Clark, A.G., Weiss, K.M., Nickerson, D.A., Taylor, S.L., Stengard, J.H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. and Sing, C.F. (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.*, **67**, 881–900.
22. Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C. *et al.* (2000) The common PPARγ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.*, **26**, 76–80.
23. Bell, G.I., Horita, S. and Karam, J.H. (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes*, **33**, 176–183.
24. Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
25. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H. *et al.* (2001) A frameshift mutation in NOD2 is associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.
26. Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T.H., Mashima, H., Schwarz, P.E. *et al.* (2000) Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.*, **26**, 163–175.
27. Neel, J.V. (1962) Diabetes mellitus: a 'thrifty' genotype rendered detrimental by progress? *Am. J. Hum. Genet.*, **14**, 353–362.
28. Slatkin, M. and Rannala, B. (1997) The sampling distribution of disease-associated alleles. *Genetics*, **147**, 1855–1861.
29. Sherry, S.T., Rogers, A.R., Harpending, H., Soodyall, H., Jenkins, T. and Stoneking, M. (1994) Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.*, **66**, 761–775.
30. Kimmel, M., Chakraborty, R., King, J.P., Bamshad, M., Watkins, W.S. and Jorde, L.B. (1998) Signatures of population expansion in microsatellite repeat data. *Genetics*, **148**, 1921–1930.
31. Reich, D.E. and Goldstein, D.B. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl Acad. Sci. USA*, **95**, 8119–8123.
32. Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A. and Feldman, M.W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.*, **16**, 1791–1798.
33. Wall, J.D. and Przeworski, M. (2000) When did the human population start increasing? *Genetics*, **155**, 1865–1874.
34. Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, **69**, 1–14.
35. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
36. Ewens, W.J. (1979) *Mathematical Population Genetics*. Springer-Verlag, New York.
37. Wright, S. (1949) Adaptation and selection. In Jepson, G.I., Simpson, G.G. and Mayr, E. (eds), *Genetics, Palaeontology and Evolution*. Princeton University Press, Princeton, pp. 365–389.
38. Slatkin, M. and Bertorelle, G. (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. *Genetics*, **158**, 865–874.
39. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**, 506–513.
40. Cox, N.J., Wapelhorst, B., Morrison, V.A., Johnson, L., Pinchuk, L., Spielman, R.S., Todd, J.A. and Concannon, P. (2001) Seven regions of the genome show evidence of linkage to type 1 diabetes in a consensus analysis of 767 multiplex families. *Am. J. Hum. Genet.*, **69**, 820–830.
41. Hanis, C.L., Boerwinkle, E., Chakraborty, R., Ellsworth, D.L., Concannon, P., Stirling, B., Morrison, V.A., Wapelhorst, B., Spielman, R.S., Gogolin-Ewens, K.J. *et al.* A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nat. Genet.*, **13**, 161–166.